知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

档案信息化的大数据问题与解决对策探析

◎ 杜晓艳

深圳大学图书馆 深圳 518060

摘要：[目的/意义]探讨档案信息化管理面临的数据量急剧增加、类型与结构日益多元复杂的现实问题。[方法/过程]结合档案的基本属性与特征，分析数字化档案在存储、利用等典型环节所具有的大数据特征，研究并阐述档案信息化过程中，新的大数据技术对数字档案存储与利用、知识发现过程的支持与应用。[结果/结论]现代大数据处理技术不仅为档案信息化管理带来一定的解决对策，同时可以促进其理论与实践的发展。

关键词：数字档案　大数据　档案信息化　存储与利用　知识发现

分类号：G271

引用格式：杜晓艳 . 档案信息化的大数据问题与解决对策探析 [J/OL]. 知识管理论坛 , 2017, 2(3): 244-249[引用日期]. http://www.kmf.ac.cn/p/1/123/.

1 引言

随着信息时代的快速发展，档案的信息化建设是大势所趋。档案信息化是档案管理从传统实体服务转向数字化信息服务模式的转变，通过数字化档案信息资源和网络化档案的管理过程实现对档案信息资源的合理管理和有效利用[1]。在大数据时代潮流下，大数据所具有的海量（Volume）、多样（Variety）、高速（Velocity）、可用与可信（Veracity）即 4V 特性，已经体现到档案信息化建设中，出现了“档案大数据”的概念[2]及在大数据技术支持下对数字化档案的深度挖掘策略[3]。然而，传统的档案管理系统难以动态扩展，越来越吃力[4]，网络化档案的管理过程迫在眉睫。特别是数字化档案信息资源本身日益成为繁重、冗长而效益低下的工作，数字化后的资源仍然存在“信息孤岛”现象而得不到有效利用。传统的管理与技术体系已经逐渐不能满足要求，如何与大数据环境和技术接轨是档案信息化面临的挑战与机遇。

2 档案信息化研究现状

国内关于档案信息化的研究最早开始于 20 世纪 90 年代末，档案信息化的研究源于社会信息时代的到来。随着时代进步和研究的不断深入，发表论文的数量呈逐年上升趋势，档案信息化逐渐成为档案学术界的研究热点。研究初期，学者们较多地关注档案信息化的相关理论研究，研究范围主要集中在档案信息化的来源，档案信息化与相关概念、相关工作之间的关系，档案信息化建设的相关内容研究，档案

作者简介：杜晓艳（ORCID:0000-0002-7614-1673），馆员，E-mail: dxiaoyan@szu.edu.cn。

收稿日期：2017-04-18　　发表日期：2017-06-26　　本文责任编辑：王善军

知识管理论坛，2017(3):244–249
DOI：10.13266/j.issn.2095–5472.2017.026

信息化过程中存在的问题及对策等方面。史丽萍[5]认为档案馆与社会信息化紧密相关，探讨了档案信息化的形成，并对未来发展趋势进行分析。李治金[6]分析了档案信息化与企业信息化之间的联系，说明信息化对企业档案事业发展的重要性。张锐[7]对档案信息化理论体系建设的有利时机、建设现状与存在问题，及完善档案信息化理论体系建设的策略和措施进行了探讨。丁立新[8]在分析我国档案信息化发展的机遇与困惑基础上，对档案信息化工作模式、应用系统建设及其运行维护的发展方向进行了趋势预测。王美琴[9]则基于我国档案信息化建设基本现状分析，指出档案信息化过程中存在的主要问题，提出加快实施档案信息化的措施。

随着物联网的出现和云计算、大数据等信息技术的兴起，社会信息化水平越来越高，同时人们对档案信息化的要求也逐渐提高。自2011年以来，国内掀起了大数据研究热潮，研究文献数量呈逐年上升趋势，大数据和档案信息化的结合也日渐紧密。我国学者围绕大数据背景，展开了一系列针对档案信息化的研究。张英奎[10]等分析了大数据时代企业档案管理所面临的主要问题，为使档案管理模式更好契合时代发展，提出了相关策略。刘国华等[11]建议从服务观念、档案信息质量、档案资源云平台构建三个方面融入并强化大数据技术应用。

国内学者同时还关注大数据技术背景下我国高校档案信息化发展与应用问题。陈晨[12]分析了高校图书馆的档案大数据及信息化现状，从软硬件基础设施、管理人员构成及其业务和安保意识、管理制度等方面提出了相应改进对策。目前国内档案信息化研究发展迅速，已经覆盖数字档案管理、档案信息服务、档案数据挖掘等内容。但总体上讲，我国档案信息化研究还处于初步发展阶段，更多地关注信息技术在档案信息化过程中的应用，理论体系尚不完善；对大数据背景下档案信息化所面临的理论基础和技术问题认知尚浅，缺乏对档案信息化技术的具体说明，研究范围和领域有待拓展，研究层次有待进一步深化。

3 档案信息化建设面临的大数据问题

当前大数据技术的示范应用主要包括社交媒体数据分析、互联网广告、地理坐标及商务智能[13]，主要应用于数据挖掘与决策层面。但从长远看，以上领域会拓展到数据长期保存、信息系统管理等方面。任何新技术的应用都与社会环境密切相关，显示着技术的社会性，并受到各种社会条件的制约和影响[14]。在档案的信息化建设过程中，大数据技术应用在元数据与数字化档案信息资源的存储、可追溯、利用的时效性、知识服务的可用性等方面面临一些问题。

3.1 数字档案资源存储问题

只有对所收藏的数字档案资源进行可信的、长久保存的系统才能称之为数字档案馆系统[15]。档案的特性在于持续记录不断发展的历史过程。因此，数字档案资源的存储是个永久的量增过程，需要不断扩充存储载体来支持这样的增长。此外，任何材质的载体受到自然环境及技术进步的影响，都会逐渐丧失载体的功能，从而影响到所记录的信息。实现永久保存就意味着按照一定的时间周期，对于需要永久保存的档案资源定期进行数据迁移，以便对数据进行载体更新、技术更新、管理更新，使得信息资源能够不断保存下去，这是档案实现可靠长期保存的基本要求。

数字档案的数据由描述档案实体内容的数据与描述数据的数据（元数据）两大部分构成。档案数据一般具有只读特性。因此，数字档案的著录、标引、索引、目录等元数据相对容易进行标准化。但是其实体数据的数据类型、格式、结构等会随着技术革新不断发生变化，其所承载的信息完整性与可靠性受到挑战。对于数据量的规模增加可以通过不断增加相应的软硬件设备来应对，但量的规模到一定程度后有可能导致对数据的控制力下降甚至系统崩溃。大数据环境下，数据迁移是最难应对的考验。

尽管可以通过传统关系数据库的三级模式来维持迁移过程软件的独立性，但迁移的数据量会越来越大；迁移数据的数据结构及附载其上的信息含义越来越复杂；迁移的周期随着技术革新节奏的加快，周期越来越短。传统的数据库模式已经不能有效应付迁移的复杂性，特别是现有的系统经过技术或管理革新重组后，数据的类型、结构、约束等都存在转型问题。维护档案的真实性和可靠性面临着巨大挑战。

3.2 数字档案资源的可追溯问题

从纵向角度通过档案能了解其反映出的基本语义、产生的背景、来源及原来制档机关的目的，而且也能够发现不同档案资料存在的相关性，即档案具有可追溯性。虽然档案本身一般是按照一事一案以案卷、全宗等作为关联的单位保存的，但是一因多果或一果多因在现实的社会环境中广泛存在。所以对档案的可追溯性并非局限在案卷内或全宗内，往往需要利用数字档案的特点进行复杂的关联查询与分析利用。而且，随着不同行业、专业领域之间的互相渗透，互相之间的相关性会越来越多，越来越复杂。此外，数字档案全宗及案卷内往往存在文本、图像、视频等异类及同类但异构的数据，用户的追溯需求也会越来越多元化。

由此使得数字档案的可追溯性在不远的将来日益成为一项艰巨的任务。即使数字化的信息系统在理论与实践方面能够在逻辑上实现这样的复杂关联，但是所导致的时间与空间复杂度会使成本巨大。此外，数字档案由于对各层次软硬件环境及原始档案管理制度的依赖，需要大量的元数据来描述，而元数据与档案内容之间虽然存在逻辑关联，但是在物理上常是独立的，这种关联往往随着技术环境的变化表现为一定的脆弱性。传统意义上的量或规模已经不再是衡量复杂性的第一要素，复杂关联与聚集引发的数据复杂性远远超过规模的复杂性效应 [16]。可追溯性是数字档案长期保存的可用性基本要求，日益复杂的关联性与高效、可用及可信是矛盾统一体，也是大数据环境下必须面对的问题。

3.3 数字档案利用的时效性问题

在一般性事务查询利用方面，对于以关系模型存储的档案元数据，标准 SQL 查询的结果与响应时间（时间复杂度）受到数据量与关联数的限制，理论与实践上不可能无限制扩大。如果数据库中包含了图像、大文本、视频等大二进制字段，检索效率更会大打折扣。此外，为了加强对数字化档案的利用，会在原始分类的基础上要求有更多的逻辑分类，以便于进行关联分析。由此，在检索过程中，会造成数据库之间、数据表之间复杂的、大数据量的关联运算。另外，现有的数字档案系统一般均要求支持全文查询，现有技术针对全文查询一般是建立在对相关文件穷举式扫描基础上的，在具体文件不确定的情况下，如果涉及到跨库、跨文件查询，在 EB 级数据量下，这几乎是不可能实现的。

因此，一般的解决的方法就是纵向不断增加层级及横向采用更广泛的分布系统，但不会解决时间复杂度越来越大、系统熵越来越大的根本问题。

3.4 数字档案知识服务的可用性问题

在对信息资源进行分析或进行知识发现研究时，首先要求信息资源能够按照知识发现主题的需要建立多维度分析模型，建立各种复杂关联。现有的数字化档案体系一般是传统纸制档案的数字化转换。受制于其传统载体及立档单位，其数字化副本在物理与逻辑结构上都存在小集中、大分散的现象。小集中指的是档案的保存逻辑上体现的立档单位一般以全宗为单位，事由以案卷为单位，关联方式一是通过文件物理存储的集中性来体现，另外通过大量的元数据描述在逻辑结构上体现；大分散指的是不同地域、不同机构之间的数字档案资料缺乏关联，形成一定规模的信息孤岛。虽然有利于保证档案案卷的整体性及体现原来制档机构的目的，但是不利于按照一定分类主题形成大规模的数据集市或数据仓库。数据挖掘形成的语义关联或知识图谱可信程度大打折扣。检索

知识管理论坛，2017(3):244–249
DOI：10.13266/j.issn.2095–5472.2017.026

查询及查询后基于批处理的分析计算在数据量及非结构化达到一定程度后，很难保证其可用性，更不能保证其高效性。

此外，档案信息化还面临着元数据与数据结构问题。现有的元数据主要存在于关系数据库中，关系结构以行记录为单位，而大数据技术环境下的数据库往往是以列为单位，这样就需要对原有的元数据结构进行重新设计，也就是现有的元数据结构也需要发生相应的变化。大数据技术应用面临着与原有系统冲突的问题。

另一方面，大数据技术应用还存在对关联粒度及层次结构制约的问题。在实际应用中，数字档案之间的逻辑关系相对具有较多的层次结构。除了档案实体文件内部的相关性外，还存在案卷与案卷之间的联系，同一案卷中"件"与"件"之间的联系，不同级别的档案管理联系。这些关联具有一定的"立体"特征。但在现有的大数据技术环境下，由于数据结构相对简单，重在对异构、海量数据的"平面"关联分析，因此，如何将数字档案的数据结构重新组织，在不破坏其固有的立体联系情况下，实现高效率的大数据分析将是极大的挑战。

4 档案信息化建设中大数据问题的解决对策

4.1 加强数字档案资源存储

档案数字化是借助计算机网络技术和多媒体技术发展而产生的一种新型档案信息形态，将各种传统载体的馆藏档案资源转化为数字化档案信息，以数字化形式存储、网络化形式传输并利用计算机系统进行管理，进而实现档案信息的快捷利用与共享[17]。数字档案数据在保存中需要按照时间序列或事由进行分类与关联，追求的目标是将存储管理由载体控制转化为软件控制。传统的模式对数据的结构、操作及约束有一定的范式要求，采用转储方式或基于分布式数据库系统的模式。通过中心管理服务器将分布在不同节点数据库中的数据实现逻辑上的统一管理，存储的方法一般是将结构化的关系模型作为元数据信息存储的数据结构，以此来关联实体档案。关系型数据库虽然能够实现比较复杂的关联，但对数据量非常敏感，具有较大的时间与空间复杂度。在档案信息化建设过程中，利用大数据存储技术加强数字档案信息资源存储，如通过 GFS（Google Files System）、HDFS（Hadoop Files System） 等分布式文件存储系统，能够处理非结构化数据并实现关联，自动建立基本的索引元数据，适合半结构化数字档案信息资源的存储与处理。

4.2 维护档案静态特征及迁移过程的可靠性

原始记录性是档案的本质属性之一，客观上要求其所依赖的软硬件环境、依附的载体及其语义能够维护其所记录信息的原始性、真实性、可靠性等静态特征，同时要求随着信息技术的发展能够实现一致性的数据迁移，从而保证档案信息的可追溯性。

大数据分布式文件存储系统能够将文件或文件夹中的对象直接转化为二进制数据序列，忽视其中的具体格式或结构，对各种形式存在的档案资源在底层实现智能化存储与处理，在更高的层次上再进行利用分析；此外，大数据技术更适应对大文件的处理，如HDFS文件系统，可将要存储的非结构化数据按照统一二进制大小（64M）进行分片、多点备份、并行处理，形成一系列的 (key,value) 键值对，然后按照键进行归并，对相同键的值进行结果汇总与合并。这也符合档案文件的组织特点（以"件"或"卷"组织成复合文件）。由此能够很地维持档案资源存储与利用过程中的完整性、可靠性，实现档案数据变换、整合及利用的智能化，可以针对档案案卷的组织特性，将其以复合文件或文件类集合的模式进行多种形式的组织，然后按照全宗建立群节点，从而简化数字档案文件存储管理的层次级别。

4.3 维护数字档案的时效性和可用性

大数据技术可以通过弱化关系降低数据模型的复杂性，统一电子文件的物理与逻辑集成（集成指在文档管理范畴内，将电子文件及其

内容信息、结构信息、背景信息采用一定标准、规范和编码进行融合 [18]）。分布式键值对的存储系统能够实现面向列的、可伸缩的数据存储模式，将不同类型、不同结构的海量数据按照列簇存储到同一文件中并实现性能良好的随机访问，使数字档案按照事由进行物理封装成为可能。相应地也可以使内容信息、结构信息、背景信息具有逻辑与物理上的统一标识与封装。此外，存储的结构支持多维特性，能够在结构上实现动态改变，可以在不影响原有数字档案内容及结构的前提下，实现行、列、时间戳的动态扩展，由此可以实现数字档案内容的动态扩展。自动生成索引的机制可将非结构化的数字档案实现半结构化，实现更紧密的结合，进一步维持数字档案在长期保存过程中的完整性。如果能够与现有系统中数字档案的标准元数据进行关联，共同实现对档案内容信息的索引及描述，会极大增强数字档案的可用性。

档案一旦形成后在内容上就不能再修改，大数据技术对数据修改的敏感或不支持并不影响档案的长期保存，反而成了档案长期存储的一种优势。首先是大数据技术所支持的文件系统通过不断增加硬盘数量实现容量的智能化增长，存储采用集群架构的管理与多重备份并基于智能化容错，读写模式采用基于二进制的分块、并行处理、合并的方式，而且所使用的文件系统一般不限制文件大小及格式。因此，在构建数字档案存储系统时，无论从逻辑上还是物理上都能够有效维护档案资源的原始记录性特征。

4.4 实现档案信息的关联性分析及知识发现

数字档案的存储在数据结构上大都具有半结构化特性。一方面由于档案数量与种类的多元导致的海量、异构等非结构化特征；另一方面数字档案一般都有结构化、标准化的元数据描述及电子标引等元素。因此，随着数字档案资源的不断增加，完全结构化或完全非结构化的资源形态均不多见。将结构化元数据与非结构化的档案实体描述数据按照一定的模式关联成半结构化模式，日渐成为数字档案资源组织的基本模式。大数据技术环境可以提供基于键值对的分布式存储与处理，能够在海量、异构数据中自动寻找出文字间的语义主题，有利于面向领域对本体的主题知识构建。此外，面向列的、可伸缩的半结构化数据库存储模式，如基于 HDFS 的 Habse 数据库管理系统，能够在行、列、时间维实现动态扩展，通过行关键字、列簇、列关键字、时间戳形成多维表。一方面能够实现复杂的半结构化与非结构化数据之间的关联，另一方面也有利于形成领域、论域、主题三个知识关联层次。由此，为异源、异构的数字档案进行数据挖掘与知识发现提供了基本的技术支持。

5 结束语

尽管档案信息化工作所面临的大数据问题突出，相关的研究重点主要聚焦在档案所具有的大数据特征及相关的管理与利用宏观策略方面 [19]，但大数据技术所具有的分布式、云计算、智能化特征，及对海量、异构数据处理所具有的优势与数字档案管理的现实需求具有一定的吻合度。与此同时，对大数据及其相关技术应用于数字档案管理的深层次问题需要进一步理解与把握，例如大数据技术如何为档案存储、迁移及跨部门与平台的知识发现提供支持，以及应用于数字档案管理所需要面对的信息系统重构、信息及数据转换层次划分问题等。大数据技术及其生态环境是信息技术发展的必然趋势，促进着通常的数字档案以及包括数字图书资源等泛化“数字档案”相关管理理论与技术的快速发展 [20]。

参考文献：

[1] 王学平 . 浅议我国档案数字化建设实践与发展策略 [J]. 档案学通讯 , 2011(6): 54-57.
[2] 鲁德武 . 试述档案大数据的定义、特征及核心内容 [J]. 档案 , 2014(4): 13-15.
[3] 张文元 , 张倩 . 大数据技术与档案数据挖掘 [J]. 档案管理 , 2016(2): 33-35.
[4] WANG X C, DING J Y. On innovation of archive

management in big data era[C]// 中国科学技术信息研究所 . 大数据时代的科技资源共享 : COINFO 2013 论文集 . 北京 : 科学出版社 , 2013: 66-72.
[5] 史丽萍 . 档案信息化形成与发展趋势 [J]. 黑龙江档案 , 1999(6): 31.
[6] 李治金 . 谈档案信息化与企业信息化 [J]. 档案与建设 , 2003(1): 52-53.
[7] 张锐 . 档案信息化理论体系建设的理性思考 [J]. 档案学研究 , 2008(2): 49-52.
[8] 丁立新 . 档案信息化的发展趋势 [J]. 档案学研究 , 2009(4): 12-14.
[9] 王美琴 . 我国档案信息化过程中的主要问题及对策 [J]. 档案学研究 , 2011(1): 64-66.
[10] 张英奎 , 王飞 , 房彦君 . 大数据时代的企业档案信息化建设 [J]. 北京工业大学学报 (社会科学版), 2014(3): 32-36.
[11] 刘国华 , 李泽锋 . 档案工作中大数据框架构建及应用思考 [J]. 档案管理 , 2014(2): 32-34.
[12] 陈晨 . 基于大数据的高校图书馆档案信息化建设探究 [J]. 兰台世界 , 2015(23): 143-144.
[13] 李战怀 , 王国仁 , 周傲英 . 从数据库视角解读大数据的研究进展与趋势 [J]. 计算机工程与科学 , 2013, 35(10): 1-11.
[14] 常立农 . 技术哲学 [M]. 长沙 : 湖南大学出版社 , 2003: 14-15.
[15] 冯惠玲 , 刘越男 . 电子文件管理国家战略 [M]. 北京 : 中国人民大学出版社 , 2011: 350.
[16] 何非 , 何克清 . 大数据及其科学问题与方法的探讨 [J]. 武汉大学学报 (理学版), 2014, 60(1): 1-12.
[17] 王美琴 . 我国档案信息化过程中的主要问题及对策 [J]. 档案学研究 , 2011(1): 64-66.
[18] 赵屹 . 基于前端控制思想的电子文件形成过程研究 [J]. 档案学研究 , 2012(3): 16-23.
[19] 石俊峰 , 周俐霞 , 付双双 . 大数据时代数字档案资源管理研究现状与趋势分析 [J]. 信息安全与通信保密 , 2014(9): 87-89.
[20] 苏新宁 . 大数据时代数字图书馆面临的机遇和挑战 [J]. 中国图书馆学报 , 2015(6): 4-12.

Analysis on Big Data Problems and Technique Supports of Archives Informatization

Du Xiaoyan
Shenzhen University Library, Shenzhen 518060

Abstract: [Purpose/significance] The realistic questions of the archives informatization management are faced with data size rapidly increasing, and their types and structures more diverse and complex. **[Method/process]** Based on the essential attribute of archives in this paper, the big data characteristics of digital archives in their storage and utilization links were analyzed, and the support of new big data techniques in the course of archives informatization, and their applications to the storage and utilization of digital archives and knowledge discovery were researched. **[Result/conclusion]** Modern processing technology for big data would not only bring certain supports for the management of archives informatization, but also promote the development of its theory and practice.

Keywords: digital archives big data archives informatization storage and utilization knowledge discovery